Re-thinking Model Inversion Attacks Against Deep Neural Networks

Ngoc-Bao Nguyen(*) Keshigeyan Chandrasegaran(*) Milad Abdollahzadeh Ngai-Man Cheung

Singapore University of Technology and Design (SUTD)

Poster **WED-PM-384**

Wed 21 Jun 4:30 p.m. PDT — 6 p.m. PDT
West Building Exhibit Halls ABC 384

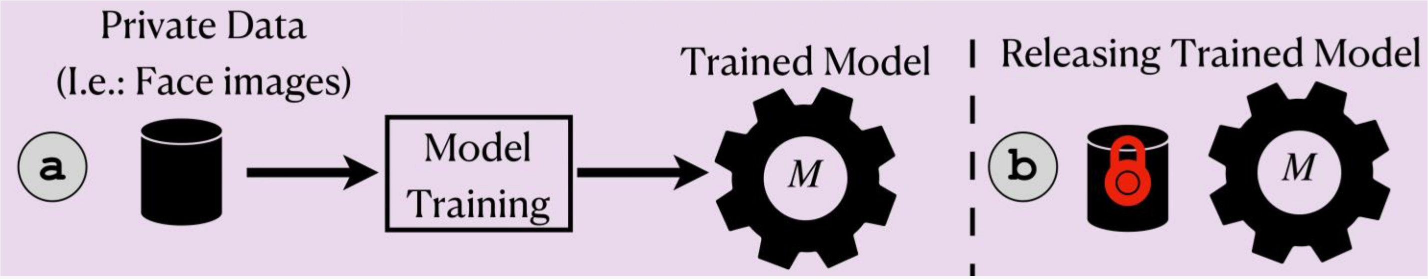
(*) Equal Contribution

Re-thinking Model Inversion Attacks Against Deep Neural Networks

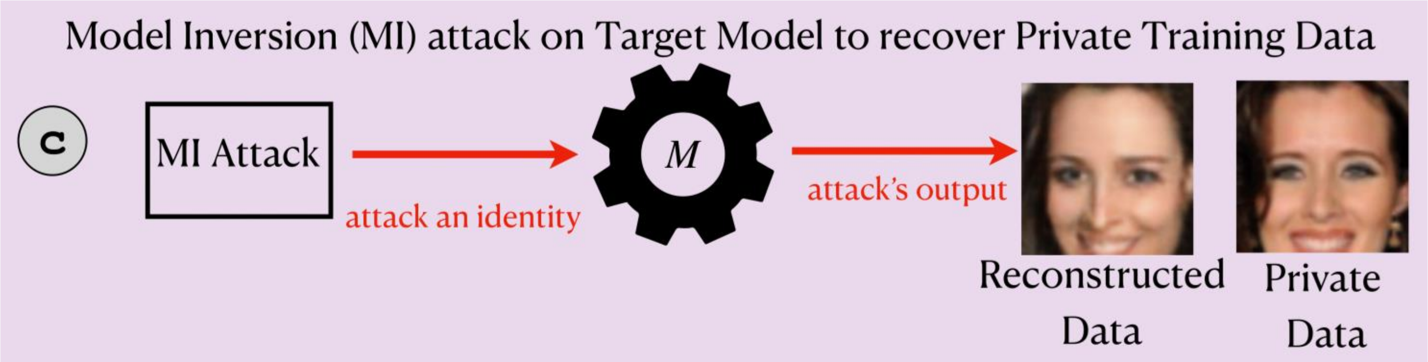
- We study Model Inversion (MI), a type of attack that aims to infer and reconstruct private training data by abusing access to a trained model.
- We analyze two fundamental issues pertaining to all state-of-the-art (SOTA) MI algorithms and propose solutions to these issues, which lead to a significant boost in performance for all SOTA MI methods.

Our results highlight the rising threats posed by MI
and prompt serious consideration regarding the privacy of machine learning.

Model Inversion (MI)



Model inversion (MI) attacks aim to infer and reconstruct private training data by abusing access to a model.



Is there a risk of data leakage for private training data when attackers abuse access to a trained model?

Prior works

Problem setup. An attacker abuses access to a model M trained on a private dataset $\mathcal{D}_{priv}$

Goal. Infer and reconstruct information about private samples in $\mathcal{D}_{priv}$

Given a desired class/ identity y, MI attacks [1,2] perform the following optimization:

$$q^*(z) = \arg\min_{q(z)} \mathbb{E}_{z \sim q(z)} \{ \underbrace{\mathcal{L}_{id}(z; y, M)}_{\text{Identity loss}} + \underbrace{\lambda \mathcal{L}_{prior}(z)}_{\text{Prior loss}} \} \quad (1)$$

The reconstructed images: $$x = G(z) \quad (2)$$

where $z \sim q^*(z)$, generator G is trained using a public dataset $\mathcal{D}_{pub}$

[1] The secret revealer: Generative model inversion attacks against deep neural networks. CVPR 2020
[2] Knowledge-enriched distributional model inversion attacks. CVPR 2021.

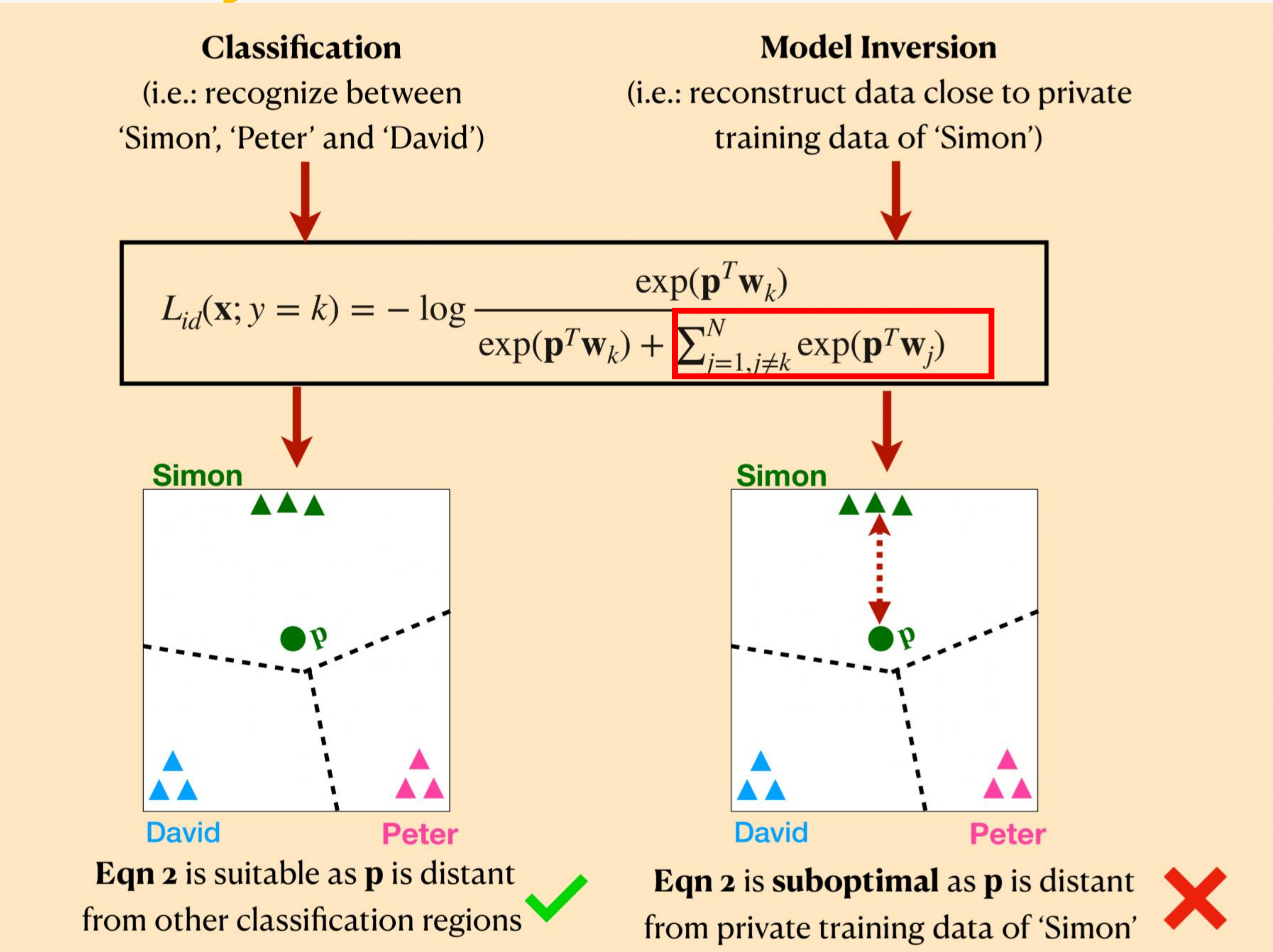
Existing Identity loss

$$\mathcal{L}_{id}(z; y = k, M) = -\log \mathbb{P}_M(y = k|G(z)) \tag{3}$$

$$= -\log \frac{\exp(p^T w_k)}{\exp(p^T w_k) + \sum_{j=1, j\neq k}^{N} \exp(p^T w_j)} \tag{4}$$

This objective can be achieved by both maximizing $\boldsymbol{exp(p^T w_k)}$
and/or minimizing $\boldsymbol{\sum_{j=1,j\neq k}^{N} exp(p^T w_j)}$

Existing Identity loss



Our proposed solution: Logit maximization (LOM)

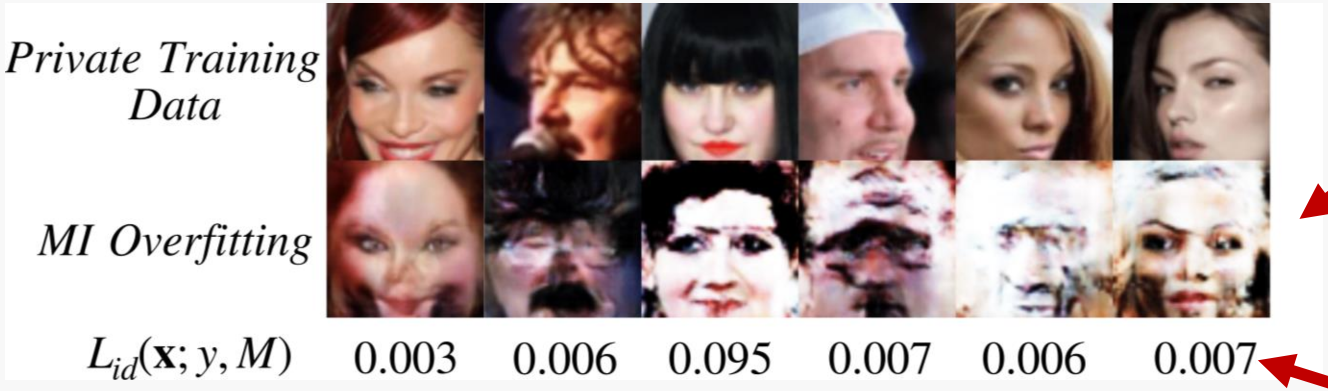
An improved formulation of MI Identity loss.
We propose to directly maximize the logit, $\boldsymbol{p^T w_k}$, instead of maximizing the log likelihood of class k for MI:

$$L_{id}^{logit}(x; y = k) = -\log p^T w_k + \lambda ||p - p_{reg}||_2^2 \quad (5)$$

where p refers to penultimate layer activations for sample x
w_i refers to the last layer weights for the i^{th} class
p_{reg} is used for regularizing p

MI overfitting

MI overfitting.
Given the fixed (target) model and the goal of MI is to reconstruct private training samples, we define MI overfitting as instances which during model inversion, the reconstructed samples fit too closely to the target model and adapt to the random variation and noise of the target model parameters, failing to adequately learn semantics of the identity.

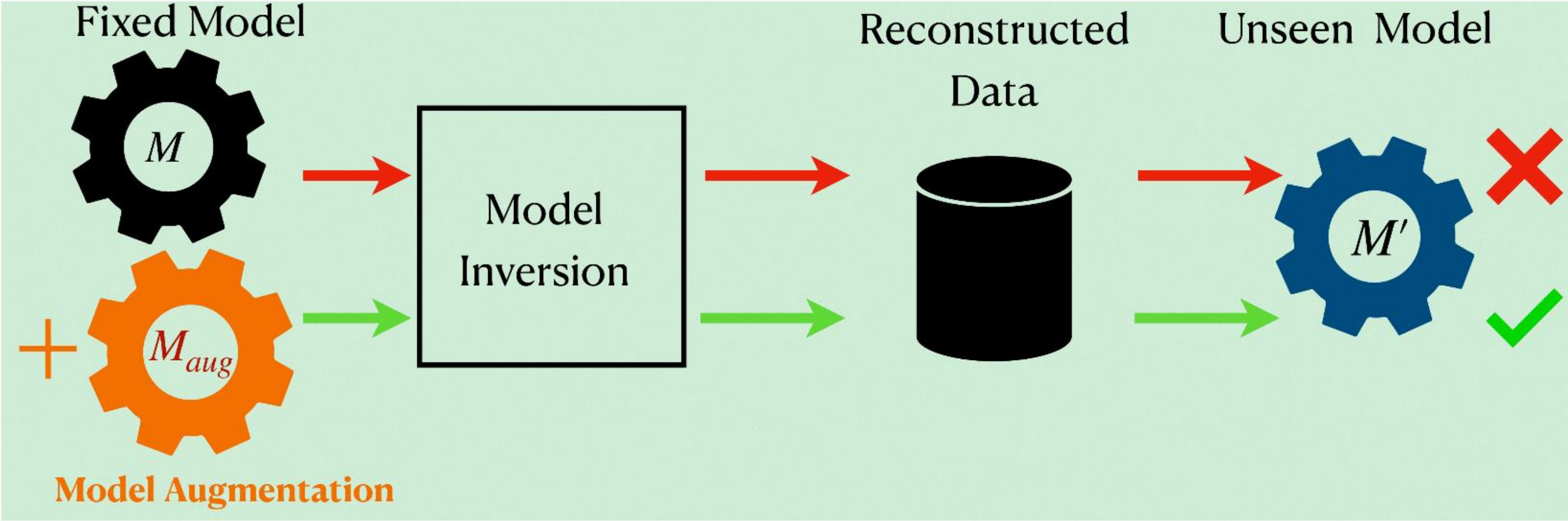


The images does not contain semantics of the identity

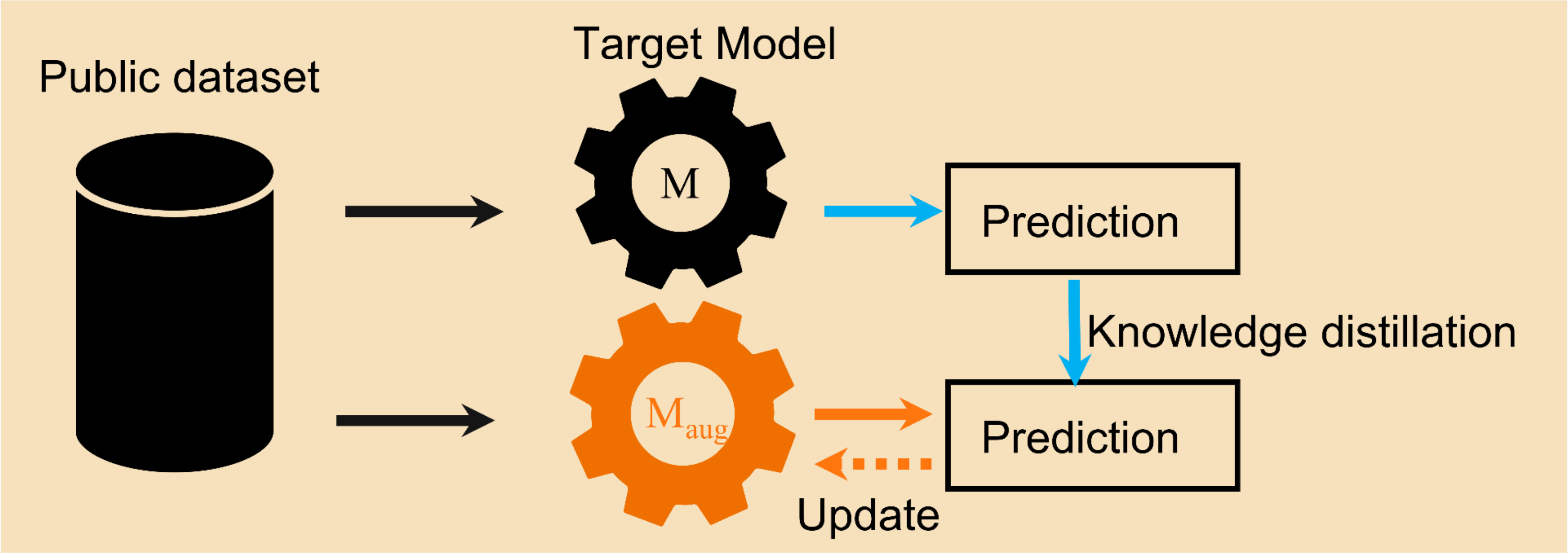
The identity loss is very small

MI overfitting in SOTA MI attacks

Our proposed solution: Model augmentation (MA)



Our proposed solution: Model augmentation (MA)



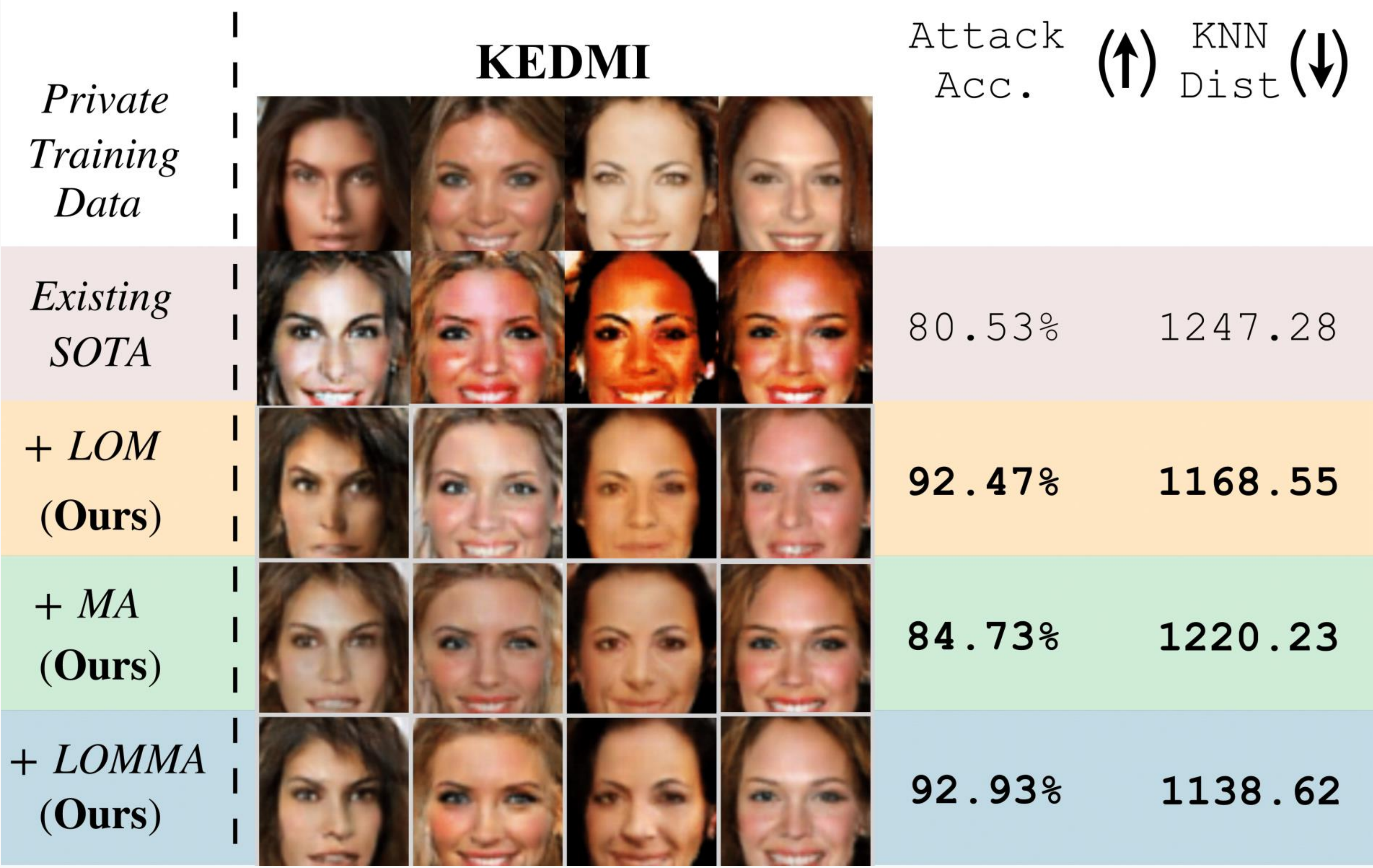
Experiments

MI attack results using different target models and public dataset

Method	Attack Acc ↑	Imp. ↑	KNN Dist ↓
CelebA/CelebA/IR152			
KEDMI	80.53 ± 3.86	-	1247.28
+ LOM (Ours)	92.47 ± 1.41	11.94	1168.55
+ MA (Ours)	84.73 ± 3.76	4.20	1220.23
+ LOMMA (Ours)	**92.93 ± 1.15**	**12.40**	**1138.62**
GMI	30.60 ± 6.54	-	1609.29
+ LOM (Ours)	78.53 ± 3.41	47.93	1289.62
+ MA (Ours)	61.20 ± 4.34	30.60	1389.99
+ LOMMA (Ours)	**82.40± 4.37**	**51.80**	**1254.32**
CelebA/CelebA/face.evoLve			
KEDMI	81.40 ± 3.25	-	1248.32
+ LOM (Ours)	92.53 ± 1.51	11.13	1183.76
+ MA (Ours)	85.07 ± 2.71	3.67	1222.02
+ LOMMA (Ours)	**93.20 ± 0.85**	**11.80**	**1154.32**
GMI	27.07 ± 6.72	-	1635.87
+ LOM (Ours)	61.67 ± 4.92	34.60	1405.35
+ MA (Ours)	74.13 ± 4.32	47.06	1352.25
+ LOMMA (Ours)	**82.33 ± 3.51**	**55.26**	**1257.50**

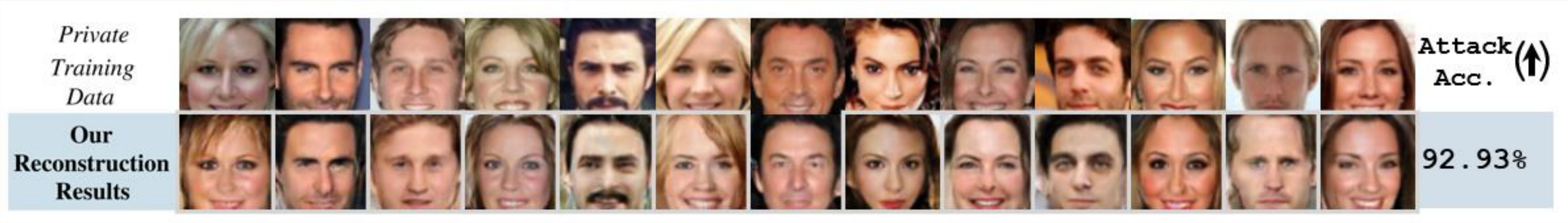
Method	Attack Acc ↑	Imp. ↑	KNN Dist ↓
CelebA/FFHQ/IR152			
KEDMI	52.87 ± 4.96	-	1418.83
+ LOM (Ours)	67.73 ± 2.29	14.86	1325.28
+ MA (Ours)	64.13 ± 4.49	11.26	1373.42
+ LOMMA (Ours)	**77.27 ± 2.01**	**24.40**	**1292.80**
GMI	17.20 ± 5.31	-	1701.76
+ LOM (Ours)	56.00 ± 5.20	38.80	1427.59
+ MA (Ours)	50.80 ± 6.89	33.60	1462.92
+ LOMMA (Ours)	**72.00 ± 6.62**	**54.80**	**1338.35**
CelebA/FFHQ/face.evoLve			
KEDMI	51.87 ± 3.88	-	1440.19
+ LOM (Ours)	69.73 ± 2.47	17.86	1379.73
+ MA (Ours)	65.73 ± 3.51	13.86	1379.09
+ LOMMA (Ours)	**73.20 ± 2.24**	**21.33**	**1321.00**
GMI	14.27 ± 4.42	-	1744.47
+ LOM (Ours)	47.93 ± 4.87	33.66	1498.19
+ MA (Ours)	46.07 ± 4.88	31.80	1500.10
+ LOMMA (Ours)	**64.33 ± 4.69**	**50.06**	**1386.33**

Experiments



Experiments

M = IR152



M = face.evoLve

Conclusion

- We analyze the existing identity loss in the SOTA and argue that it is sub-optimal for Model Inversion. Our proposed identity loss aligns better with the goal of MI.
- We formalize the new concept of MI overfitting and propose model augmentation to alleviate MI overfitting.
- Our proposed solutions are simple and easy to integrate into existing SOTA MI attacks, resulting in a significant improvement in attack accuracy.
- Our findings demonstrate a clear risk of sensitive information leakage from deep learning models.

Re-thinking Model Inversion Attacks Against Deep Neural Networks

Poster **WED-PM-384**

Wed 21 Jun 4:30 p.m. PDT — 6 p.m. PDT
West Building Exhibit Halls ABC 384

Our paper, code, pre-trained models, demo

https://ngoc-nguyen-0.github.io/re-thinking_model_inversion_attacks/